An Oracle White Paper
November 2012

Oracle Exalogic Elastic Cloud: Advanced I/O Virtualization Architecture for Consolidating High-Performance Workloads

ORACLE

Disclaimer

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Executive Overview

Oracle Exalogic Elastic Cloud is an engineered system that consists of integrated hardware and software designed, optimized, and certified for deploying Oracle business applications, Oracle Fusion Middleware, and third-party software products. Oracle Exalogic is designed to meet the highest standards of reliability, serviceability, and performance for widely varied, performance-sensitive, mission-critical workloads. Oracle Exalogic dramatically improves the performance of virtually any Oracle Linux, Oracle Solaris, or Java application with no code changes required and, compared to traditional enterprise application platforms, it reduces application implementation costs and ongoing costs while reducing deployment risk.

Oracle Exalogic has proven itself in deployments around the world, helping companies close business faster, improve their customers' online buying experience, and respond more quickly to market opportunities. While these results strongly benefit business application users and the IT teams that support them, Oracle Exalogic is also changing the way data centers operate. By eliminating costly hardware and software integration work, data center managers can greatly reduce the pain of system installations and upgrades and also simplify the management of systems in production.

Oracle Exalogic is an open system, assembled by Oracle from Oracle's portfolio of standards-based, best-of-breed component products and technologies. The Oracle Exalogic system reflects best practices learned from thousands of customer deployments and extensive laboratory testing. While the main components of the Oracle Exalogic system are individually orderable, an Oracle Exalogic configuration is more than just the sum of its parts.

In Oracle Exalogic, the design of the components themselves was influenced by the requirements of the application: Oracle has made optimizations and enhancements to Oracle Exalogic components and to Oracle's virtualization and middleware technologies that cannot be made by customers or by any third party. These range from on-chip network virtualization to operating system and Java Virtual Machine support for extremely high-performance Remote Direct Memory Access (RDMA) and Oracle Exalogic–aware workload management in Oracle's Java EE application server.

One of the key enhancements in the current generation of Oracle Exalogic systems—and the focus of this whitepaper—is Oracle's incorporation of virtualized InfiniBand I/O interconnects using Single Root I/O Virtualization (SR-IOV) technology to permit the system to share the internal InfiniBand network and storage fabric between as many as 63 virtual machines per physical server node with near-native performance simultaneously allowing both high performance and high workload consolidation.

With the latest version of the Exalogic Elastic Cloud Software, Oracle delivers a combination of capabilities unique in the industry in terms of their ability to deliver extreme, industry-leading performance while still enabling high server consolidation ratios for maximum data center efficiency.

Such an advanced set of "no compromises" capabilities is difficult to impossible to create in a "do it yourself" environment with off-the-shelf software and hardware from multiple vendors where it is necessary to finely tune and tightly integrate each component to create a seamless whole. Perhaps even more challenging is maintaining such a system over time with all the various components being enhanced by multiple vendors independently, requiring a never-ending testing and benchmarking effort. Only Oracle, with its complete, sophisticated portfolio of hardware and software, has the resources to deliver such a tightly integrated solution to meet the performance, scalability, security, and efficiency needs of your business.

Introduction

The first version of Oracle Exalogic Elastic Cloud integrated industry-standard components with an advanced communication (I/O) fabric that tied all the system components together and provided the basis for Oracle Exalogic's reliability, scalability, and performance. This system fabric, known as

Exabus, delivers high application performance and exceptionally low network and storage I/O latency for high-end computing environments. The same advanced communication fabric supports direct connection to the Oracle Exadata Database Machine. Oracle Exadata provides extreme performance for both data warehousing and OLTP applications, making it the ideal platform for consolidation on private clouds.

The current release of the Exalogic Elastic Cloud Software includes a tightly integrated server virtualization layer with unique capabilities allowing the consolidation of multiple, separate virtual machines containing applications or middleware on each server node while introducing essentially no I/O virtualization overhead to the Exabus InfiniBand network and storage fabric.

The purpose of server virtualization is to fundamentally isolate the operating system and applications stack from the constraints and boundaries of the underlying physical servers. By doing this, multiple virtual machines can be presented with the impression that they are each running on their own physical hardware when, in fact, they are sharing a physical server with other virtual machines. This allows server consolidation in order to maximize the utilization of server hardware, while minimizing costs associated with the proliferation of physical servers—namely hardware, cooling, and real estate expenses.

This hardware isolation is accomplished by inserting a very thin layer of software between the OS in the virtual machine and the underlying hardware to either directly emulate the hardware or to otherwise manage the flow and control of everything from CPU scheduling across the multiple VMs, to I/O management, to error handling, and so on.

Advancements in server virtualization software, as well as in the ability of the server hardware itself to facilitate and accelerate virtualization tasks, have all but eliminated the performance impact of virtualization. However, to get the best of both worlds—high server consolidation ratios and consistently high-performance—extensive, advanced engineering and integration across all the major components, including the hardware, the virtualization software, the operating systems, and the I/O stack, is required.

Through Oracle's unique ownership of the entire application stack—from the hardware to the virtualization layer, operating system, middleware, and applications—only Oracle can engineer a complete solution to give you the best of both worlds and eliminate the need to choose between high consolidation ratios and high performance.

Hardware and Software Engineered to Work Together

The Oracle Exalogic system consists of two major elements:

- Exalogic Elastic Cloud X2-2 (and later generations): A high-performance hardware system, assembled by Oracle, that integrates storage, networking, and compute resources using the Exabus high-performance I/O backplane built on Quad Data Rate (QDR) InfiniBand technology
- Exalogic Elastic Cloud Software: An essential package of Oracle Exalogic–specific software, device drivers, and firmware that is pre-integrated with Oracle VM, Oracle Linux, and Oracle Solaris, enabling Oracle Exalogic's advanced performance, consolidation, and management features

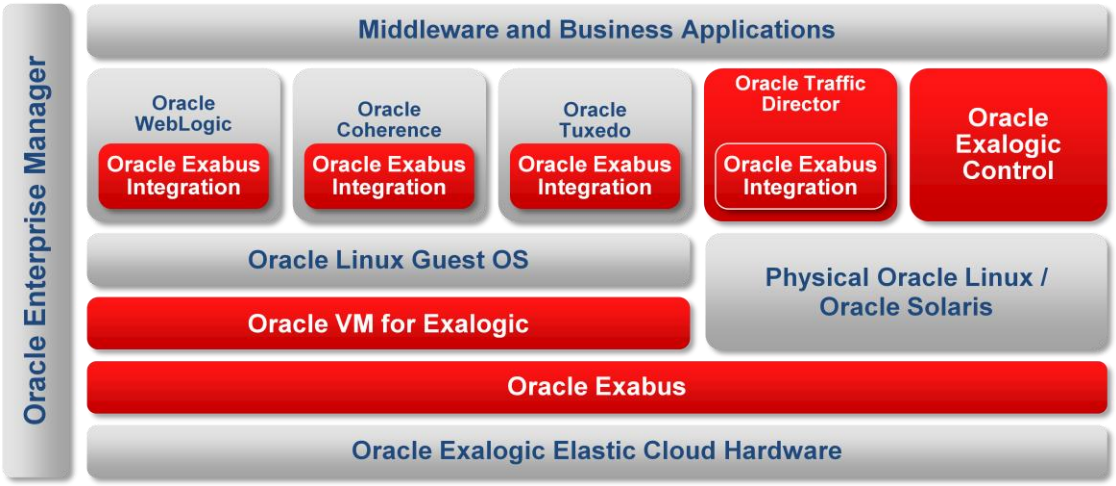

Figure 1: Oracle Exalogic system architecture.

Exabus

The defining architectural feature of Oracle Exalogic is the communication (I/O) fabric that ties all the system components together and provides the basis for Oracle Exalogic's reliability, scalability, and performance. Within Oracle Exalogic, this I/O subsystem is called Exabus, since it performs the function of extending and connecting the PCIe-based system bus used within each of the major system components. Oracle Exabus is based on Quad Data Rate (QDR) InfiniBand and consists of hardware, software, and firmware distributed throughout the system and involving every major system component.

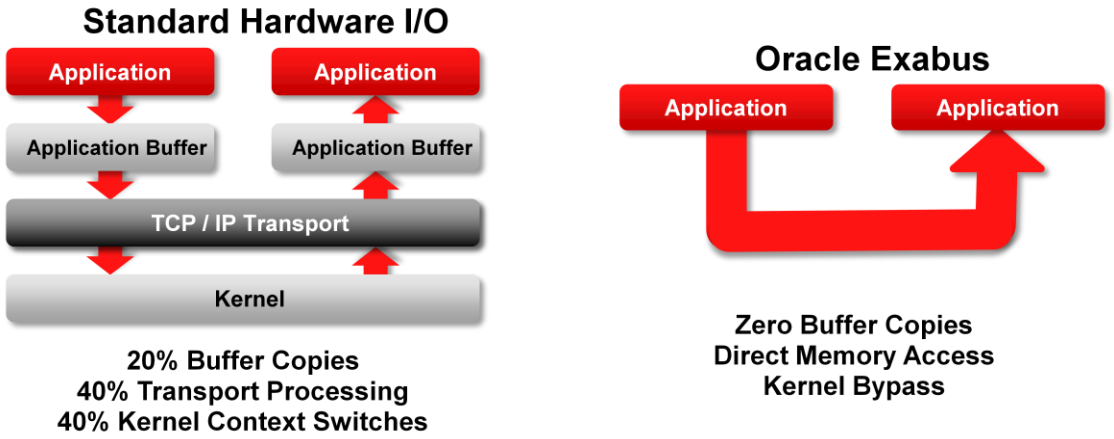

Figure 2: Exabus significantly accelerates data transfer between applications.

QDR InfiniBand was selected as the foundation technology for Oracle Exabus for several reasons:

- Oracle's InfiniBand products provide the greatest available bandwidth per physical port (40 Gb/sec) and the lowest latency (~1.07 µsec) of any standard interconnect technology available today[1] allowing applications to reclaim compute capacity that is otherwise wasted waiting on slow communication links.
- InfiniBand provides reliable delivery, security, and quality of service at the physical layer in the networking stack and it natively supports kernel bypass operations, eliminating much of the inefficiency of using system CPU and main memory.
- Oracle's InfiniBand products support upper-stack protocols such as IP over InfiniBand (IPoIB) and Ethernet over InfiniBand (EoIB), making it possible for existing applications to run without modification and still benefit from enhanced performance.

In the latest version of Oracle Exalogic, Oracle has virtualized the InfiniBand connectivity in Exabus using state-of-the-art, standards-based technology to permit the consolidation of multiple virtual machines per physical server with no impact on performance.

[1] 3.7x the throughput and 1/5 the latency of 10 GbE, the next best option, according to http://www.hpcadvisorycouncil.com/pdf/IB_and_10GigE_in_HPC.pdf8

The Challenge: High Performance *and* High Consolidation

You want the highest performance and the highest quality of service for all your mission-critical applications, but you also need to deal with the realities of today's modern data centers: You need to deal with cost and space constraints, and you need to stay agile and able to respond efficiently to both expected and unexpected changes to the business.

This reality often means you need to virtualize and deploy multiple applications per server to ensure that your assets are fully utilized and also to keep costs associated with power, cooling, and real estate to a minimum. However, the challenge can be how to achieve a high enough consolidation ratio to achieve the cost benefits you need while still being able to provide the exceptional, predictable performance required from your core applications. Oracle Exalogic Elastic Cloud has the answer.

Designed for "no compromises" performance and consolidation, the Oracle Exalogic's Exabus has been engineered to leverage a technique known as Single-Root I/O Virtualization (SR-IOV). SR-IOV eliminates virtualization overhead to deliver the maximum performance and scalability, while also allowing the same InfiniBand I/O adapter to be shared by up to 63 virtual machines, each with a redundant pair of InfiniBand connections, enabling highly efficient, consolidated operations. SR-IOV's unique ability to nearly eliminate virtualization overhead while still allowing the sharing of hardware permits a much higher server consolidation ratio and higher performance compared to other server virtualization solutions available in the market. Solutions without SR-IOV allow only *either* direct access (delivering high performance but no server consolidation) *or* device sharing (allowing VM consolidation but at the cost of reduced performance and scalability). Other solutions force you to choose between performance and consolidation; only Oracle Exalogic allows IT to efficiently deliver both the ultra-high performance that the business demands with the flexibility that the IT Operations department needs.

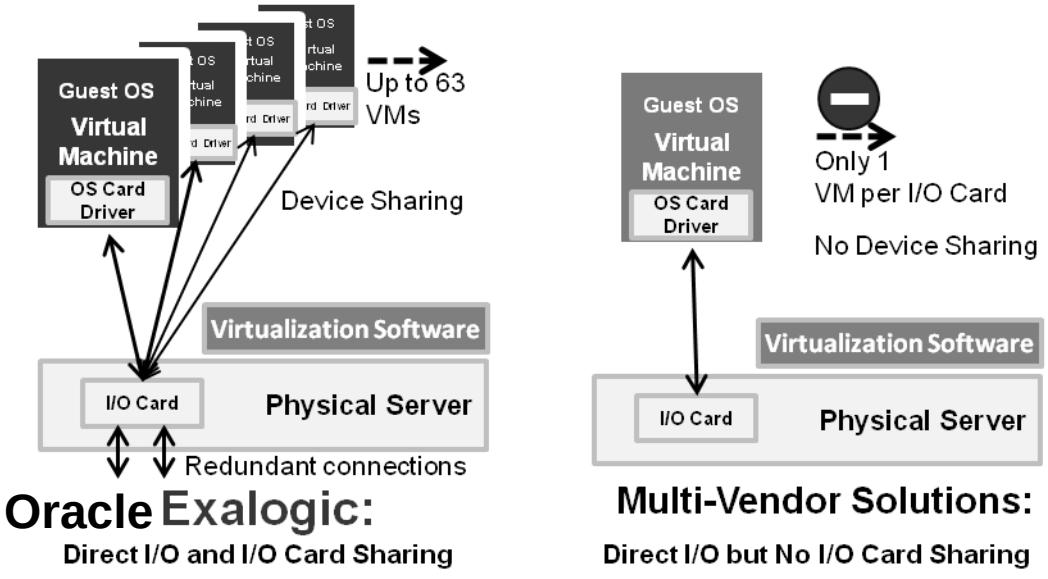

Figure 3. Oracle Exalogic compared to solutions from multiple vendors.

The Impact of Different I/O Virtualization Techniques on Performance

Today, there are three general methods of virtualizing I/O in an x86-architecture system and each method affects overall application performance differently:

- Software-based sharing
- Direct device assignment
- Single-Root I/O Virtualization (SR-IOV)

Software-Based Sharing

This is what most people think of when they think of virtualization and it is the default configuration for essentially all commercially available x86-based server virtualization products: Server virtualization software sits between the device driver software in the operating system of the virtual machine and the actual device hardware in the server. Simply put, this means the OS device driver is talking to virtualization software which, in turn, is then talking to the actual server hardware to complete the I/O operations. This approach has the benefit of allowing easy sharing of the physical hardware, because the virtualization software can allow multiple VMs to believe they all have their own set of physical hardware when, in fact, they are sharing the physical hardware, and the virtualization software does all the work to manage conflicts and ensure all the VMs run concurrently and reliably.

The downside of this approach is that the virtualization software provides its services at a performance cost: The virtualization layer requires at least some time to "traffic control" each and every I/O operation, and the more I/O operations and VMs there are, the busier the virtualization layer gets, with a potential impact on scalability under application load. As stated above, with modern virtualization software and hardware, the performance impact from overhead is increasingly very small but it can be noticeable when extreme performance is required.

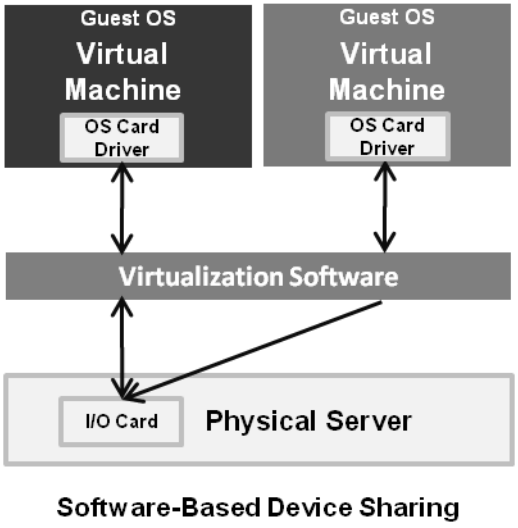

Figure 4. Software-based sharing architecture.

Direct Device Assignment

As virtualization has become a fully accepted part of production enterprise data centers, many workloads that might have been thought of as poor candidates for virtualization are now being virtualized for operational reasons. For some of these applications, where ultimate performance is essential, techniques have been developed to further reduce or even eliminate the performance and scalability risks sometimes associated with software-based sharing. One of the most common techniques for maximizing performance for virtual machines is known as direct device assignment.

Unlike software-based sharing—where each VM is talking to the virtualization layer, which is then talking to the physical hardware—direct device assignment allows the device driver in the operating system of the virtual machine to "see" and talk directly to the physical I/O hardware without having to go through the server virtualization software: The hardware device is assigned directly and exclusively to a specific, explicitly named VM (or possibly a couple of VMs, but usually only one). This effectively makes the I/O path perform and scale just as it would on bare-metal servers that aren't virtualized.

Direct device assignment eliminates the performance overhead and, thus, many scalability risks potentially associated with software-based sharing, and it can ensure more predictable and reliable I/O performance levels by essentially having I/O interfaces that are dedicated explicitly to (typically) one VM. As a result, this approach is excellent for delivering optimum performance and scalability for the one VM it is assigned to, but there are trade-offs.

The downside of direct device assignment alone is that by not requiring the I/O to go through the virtualization layer, the virtualization layer does not have the ability to deliver some of the intended benefits of virtualization. For example, direct device assignment products such as VMware's DirectPath typically permit no sharing of the I/O hardware: You might be able to configure only one VM per I/O card (in this example, an InfiniBand adapter) or, perhaps, up to one VM per physical card port. If you have a server that has only two card slots, you might be able to have only two virtual machines on that server, which might not be a cost-effective solution.

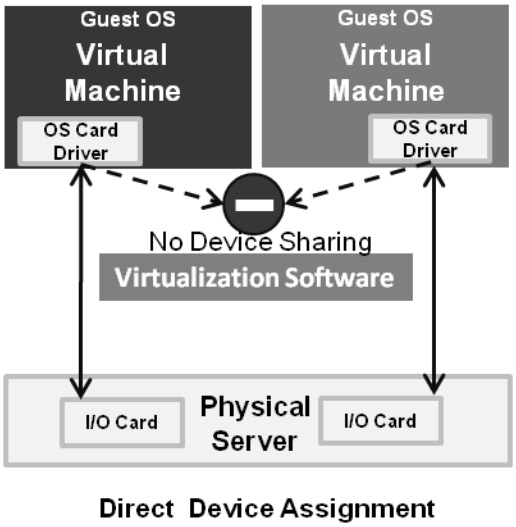

Figure 5. Direct device assignment architecture.

The ideal solution would provide a hybrid set of benefits. It would deliver the performance benefits of direct device assignment while also allowing more operational flexibility associated with full software-based hardware sharing. In order to meet these demands, SR-IOV was been developed in the industry and engineered into the Oracle Exalogic system.

Single-Root I/O Virtualization

SR-IOV is a newer industry-standard technique that delivers high performance and high scalability, while also allowing high operational efficiency through workload consolidation. With SR-IOV, and by using the latest generation of server hardware available, each individual InfiniBand I/O adapter in the Oracle Exalogic system can present itself directly to the operating system instance running in every virtual machine, but it can do so to a high number of virtual machines simultaneously rather than to just one. Oracle Exalogic supports up to 63 VMs per physical server, with each VM configured for I/O across an adapter containing a pair of InfiniBand ports for redundancy. In other words, it allows both direct device assignment as well as device sharing.

As with direct device assignment, with SR-IOV, the operating system drivers are essentially talking directly to the physical hardware interface without having to go through the server virtualization layer, which eliminates performance-sapping overhead. But since the hardware itself can present multiple instances—multiple interfaces—of itself, many VMs can share the same redundant I/O device, allowing much higher levels of server workload consolidation.

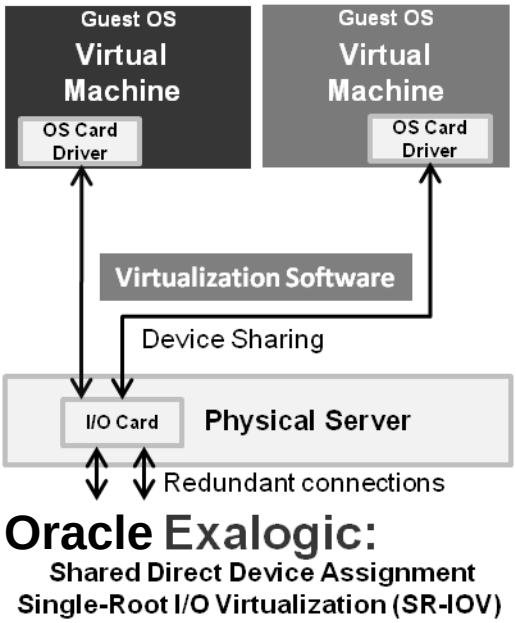

Figure 6. SR-IOV architecture.

To date, one of the challenges for virtualization solution vendors has been the sophistication that SR-IOV requires. SR-IOV requires extremely tight engineering and integration between the I/O channel adapter and BIOS, the device driver, the virtualization software, and the operating system to ensure reliable, scalable, high performance. This tight engineering and integration would be extremely difficult to deliver to users as a complete solution if all the components were delivered independently by four or five (or more) vendors, and users had to deal simultaneously with all these vendors.

Even if the solution were delivered in a manner that seemed to work well initially, maintaining such a complex environment over time would be difficult with so many independent vendors releasing updates without coordination. A new update to any component from any vendor could create significant issues throughout the solution, because users could not be sure their specific configuration had ever been tested.

Oracle eliminates these concerns by integrating the complete hardware and software stack and performing extensive testing of the virtualization solution with actual business applications being used by Oracle customers. Since all Oracle Exalogic systems are alike, Oracle can test system updates on its own systems, and these updates are guaranteed to work on all customer systems.

Full Stack, One Management Solution: Exalogic Control Software

A solution as sophisticated as Oracle Exalogic is much more than just virtualization, and that means the system management software needs to manage more than the virtualization environment. Oracle Exalogic comes with Exalogic Control software to allow users to manage Oracle Exalogic as a holistic system, including the management of the physical server , network, and storage infrastructure; the virtualization layer; operating systems; and application deployments up the stack. Rather than leaving users to take on the risk and burden of being their own multi-vendor system management integrator, Oracle delivers an out-of-box solution that takes care of complexity to enable faster, more predictable deployments as well as richer, ongoing management of the complete application stack from a single, well-integrated management solution.

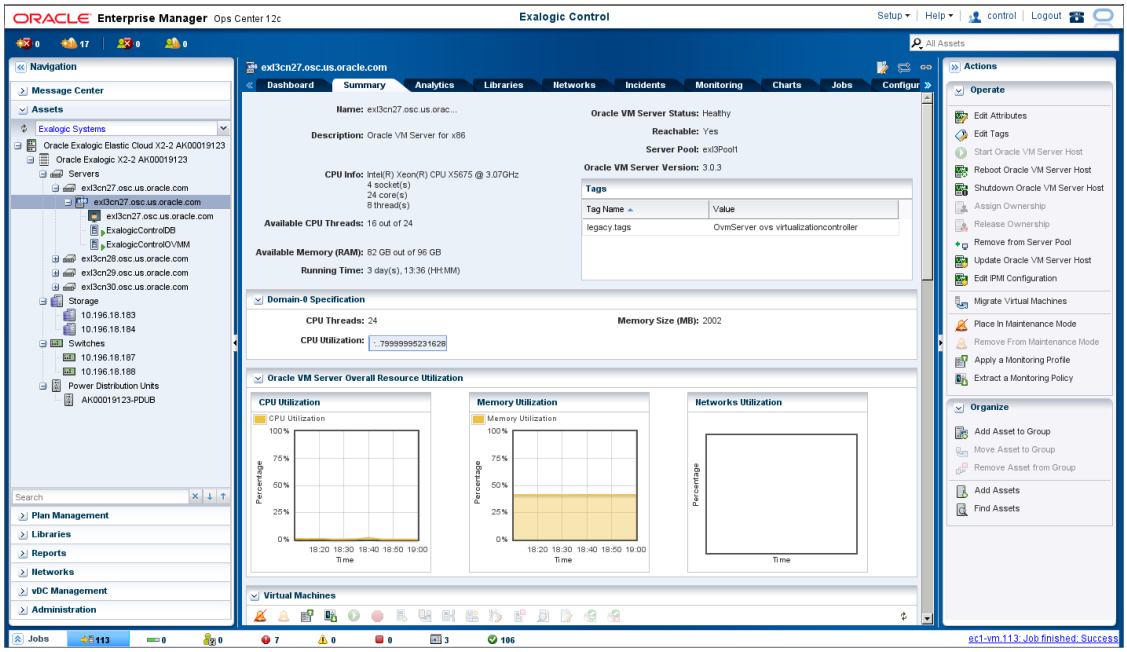

Figure 7: Exalogic Control is a single console for all hardware and software administration tasks.

Summary

Only Oracle, which has complete integration control over every aspect of the solution, can both deliver and maintain such a sophisticated, ultra-high performance solution as Oracle Exalogic, which works the first time and every time to host your most critical and highest performing enterprise applications and middleware.

Oracle Exalogic Elastic Cloud:
Advanced I/O Virtualization Architecture for
Consolidating High-Performance Workloads

October 2012
Author: Adam Hawley
Contributing Author: Yoav Eilat

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com

Oracle is committed to developing practices and products that help protect the environment

Hardware and Software, Engineered to Work Together